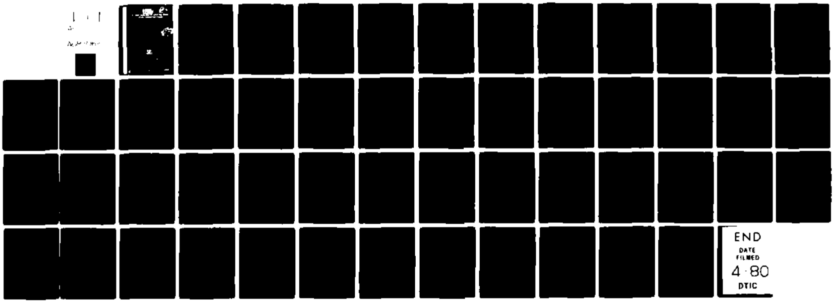
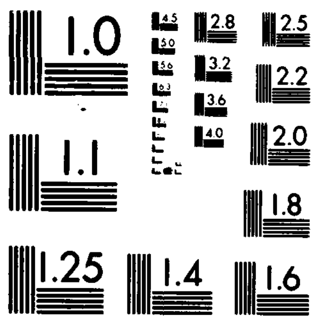
AD-A082 066 MISSOURI UNIV-COLUMBIA TAILORED TESTING RESEARCH LAB F/G 5/9
A SUCCESSFUL APPLICATION OF LATENT TRAIT THEORY TO TAILORED ACH--ETC(U)
FEB 80 R L MCKINLEY, M D RECKASE N00014-77-C-0097
UNCLASSIFIED RR-80-1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (9) Research Report [redacted]
2. GOVT ACCESSION NO. (14) RR-80-1
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) A Successful Application of Latent Trait Theory to Tailored Achievement Testing.
5. TYPE OF REPORT & PERIOD COVERED: Technical Report.
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Robert L./McKinley [redacted] Mark D./Reckase
8. CONTRACT OR GRANT NUMBER(s) (15) N00014-77-C-0097
9. PERFORMING ORGANIZATION NAME AND ADDRESS: ~~Department of Educational Psychology~~ Tailored Testing Research Lab, University of Missouri-Columbia, Columbia, MO 65211
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, VA 22217
12. REPORT DATE (11) Feb 80
13. NUMBER OF PAGES (12) 42
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
18. SUPPLEMENTARY NOTES
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Testing
Tailored Testing
Achievement Testing
Latent Trait Models
Rasch Model
Computerized Testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A live tailored achievement testing study was conducted to compare procedures based on the one- and three-parameter logistic models. Previous studies, investigating the application of these models to achievement testing, have yielded inconclusive results because of methodological problems. Close scrutiny of these investigations indicated two problems that apparently contributed to the ambiguous results. One problem was the procedures by which item calibrations were linked, and the other

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

410502

#20 (Cont.)

problem was in the item selection procedures. This second problem concerned stepsize, points of entry into the item pools, and information cutoff levels. The objective of the current study was to compare the one- and three-parameter logistic models using the improved procedures. A total of 88 students enrolled in an introductory measurement course at the University of Missouri-Columbia served as examinees for the study. A counterbalanced test-retest design was employed, in which there were two separate test sessions one week apart for each examinee. Comparisons were based upon (a) test-retest reliability, (b) ability estimates yielded by the procedures, (c) the information yielded by the procedures, (d) the number of items the methods administered, (e) goodness of fit of the models based on mean square deviations, and (f) the correlations of estimated true scores, based on ability estimates, with an outside criterion. In addition, an attitude survey was administered after each test session to determine student attitudes toward the tailored tests. The results of the study indicated that both tailored tests had higher reliabilities than a conventional paper-and-pencil test over the same material. The three-parameter procedure had higher test information than the one-parameter procedure and the conventional test. Neither procedure yielded satisfactory content validity. The attitude survey results indicated generally favorable student attitudes toward tailored testing.

CONTENTS

A Successful Application of Latent Trait Theory to Tailored Achievement Testing

Tailored testing has been proposed as an alternative measurement technique because of its potential for dealing with some of the major problems of conventional testing. Conventional testing, in which the same test items are given to all examinees, often results in test items of inappropriate difficulty being administered to many examinees. If test items are too difficult, an examinee may resort to random guessing or even omission of items, and if the items are not difficult enough, the test may not be challenging to the examinee. As a result, the standard error of measurement for conventional tests usually is higher at the extremes of the ability range, resulting in tests that are most accurate for examinees of average ability. This restricted range of accuracy is reflected in lowered test reliabilities.

Other problems, such as time limit pressures and the effects of test administration differences (Weiss, 1974), may also affect the precision of measurement of conventional tests. In order to deal with these problems, tailored testing procedures were developed (Lord, 1970). The purpose of this report is to describe a successful application of tailored testing procedures to achievement measurement. First, however, it may be helpful to discuss both the rationale and primary characteristics of tailored testing, and earlier attempts at its utilization.

Tailored testing procedures were designed to reduce the errors of measurement when estimating an examinee's ability or level of achievement by attempting to administer to each examinee only items of appropriate difficulty. This is accomplished by selecting for administration items that maximize the information about an examinee's estimated ability level. That is, each examinee receives a test which is "tailored" to his ability level. This tailoring hopefully results in increased precision of measurement.

The implementation of tailored testing procedures usually requires computer capabilities. One reason a computer is needed is that tailored testing is often based on item characteristic curve (ICC) theory (Lord, 1952; Lord and Novick, 1968). ICC theory involves mathematical models of sufficient sophistication as to require the use of a computer for parameter estimation. One of the first requirements for tailored testing is a precalibrated pool of items from which test items can be selected for administration. The calibration of the item pool is usually accomplished by using one of several existing calibration programs (Wright and Panchapakesan, 1969; Wood, Wingersky, and Lord, 1976; and Urry, 1975) on conventional test item response data in order to obtain item parameter estimates for the one-parameter or three-parameter models.

Another step which requires computer capabilities is the operation of the tailored testing procedures on an interactive basis with the examinee.

This tailored testing program is controlled by a number of program parameters, such as the point of entry into the item pool, the procedure for estimating ability (usually either a Bayesian or maximum likelihood technique), the item selection method, and a rule for terminating the test.

Once the item pool has been created and the procedures implemented, there are several problems that may arise. Among these is a possible lowering of the quality of item calibrations when it is necessary to link small sample calibrations of several tests in order to create a sufficiently large item pool. Another problem is the nonconvergence of the ability estimation procedure, and a third stems from possible violations of the assumptions of the latent trait models. This last case may occur when an extension is made from ability testing to the measurement of achievement. In the research reported here an attempt to solve these problems will be presented.

There are a number of models available for use in tailored testing, most of which belong to a class of models referred to as latent trait models. Within this class are a number of ICC models, also known as Item Response Theory (IRT) models. The particular models chosen for this study are described below.

Latent Trait Models

The Rasch (1960), or one-parameter logistic (1PL) model, as described by Wright (1977), requires one ability parameter, θ_j, for each examinee, and one difficulty parameter, b_i, for each item in order to describe the interaction of an examinee and an item. In exponentional form the 1PL model is given by

$$P(u_{ij}) = \frac{\exp(u_{ij}(\theta_j - b_i))}{1 + \exp(\theta_j - b_i)} \qquad (1)$$

where u_{ij} is the score (0 or 1) on Item i for Examinee j, θ_j and b_i are as defined above, and $P(u_{ij})$ is the probability that u_{ij} is 0 or 1.

The three-parameter logistic (3PL) model as presented by Birnbaum (1968) requires three parameters for each item. As in the 1PL model, the 3PL model requires one ability parameter for each examinee. The 3PL model is given by

$$P_i(\theta_j) = P(u_{ij} = 1) = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \qquad (2)$$

where θ_j and b_i are as defined above, a_i is the item discrimination parameter, c_i is the item guessing parameter, and D is a scaling constant equal to 1.7.

Both these models assume that the items are dichotomously scored, and that local independence holds. Also, the assumption is made that the latent trait being measured is unidimensional. (For a full discussion of the assumptions of these models see Lord and Novick, 1968.) Of particular significance is the assumption of unidimensionality. When applying factor analytic methods to ability tests, generally one dominant factor is found. But achievement tests are usually constructed with a goal of multidimensional measurement. This multidimensionality requires the serious consideration of the robustness of the models to the violation of the unidimensionality assumption when latent trait models are applied to achievement testing. Before making this examination it will be helpful to summarize the results of a previous study that used a similar tailored testing methodology and that demonstrated that tailored testing procedures could be successfully applied to a unidimensional vocabulary test (Koch and Reckase, 1978).

Vocabulary Tailored Testing Study

The purpose of the vocabulary study was to compare the 1PL and 3PL models in a tailored testing application to vocabulary ability measurement. The calibration programs used were the MAX program (Wright and Panchapakesan, 1969) for the 1PL model and the LOGIST program (Wood, Wingersky, and Lord, 1976) for the 3PL model. Items were selected to maximiz the information function (Birnbaum, 1968) for the maximum likelihood ability estimate.

The results of this study indicated that, while there were some problems, either of the two models could be successfully applied to vocabulary ability measurement. In particular, the reliabilities reported (a combination of test-retest and equivalent forms reliabilities) were $r = .77$ for the 3PL procedure and $r = .61$ for the 1PL procedure. In terms of information, the 3PL procedure outperformed the 1PL procedure, and, in the ability estimate levels between -2.0 and +.50, the 3PL procedure actually yielded greater information than the longer traditional paper-and-pencil test.

One of the problems encountered in this study was the failure of the 3PL procedure to converge to ability estimates in nearly one-third of the cases. When these cases were included in the analyses the 3PL reliability dropped to $r = .36$. The hypothesis was put forward that the cases of nonconvergence occurred because the items in the item pool were too difficult for many of the examinees.

Tailored Achievement Testing

The vocabulary test in the above study was, of course, an ability test, and relatively unidimensional (the first factor accounted for 41% of the variance). The measurement of achievement presents quite a different problem. The multidimensionality of achievement tests raises the

question of the robustness of ICC theory with respect to the violation of the unidimensionality assumption.

Very little has been published in the literature dealing with applications of tailored testing to achievement measurement. In one study conducted by Bejar, Weiss, and Kingsbury (1977), a biology achievement test was used, but that test was found to have a very dominant first factor. Not surprisingly the calibration of the item pool with the ICC model proved adequate. The use of the ICC model on a one factor achievement test would not be expected to differ much from use on a unidimensional ability test.

Research reported by Brown and Weiss (1977), in which a tailored testing procedure was used for an achievement test having several content areas, indicated that utilizing inter-subtest branching can provide precision of measurement equal to that of the conventional achievement test. However, in this study each content area was calibrated separately, rather than together as a multidimensional item pool. Therefore, even though tailored testing procedures were applied to a multidimensional achievement test, the issue of the robustness of the ICC model with respect to violation of the assumption of unidimensionality was not addressed.

The issue was addressed, however, in a study reported by Koch and Reckase (1979). In this study achievement tests were not calibrated by content area, but rather each test was calibrated as a whole. The achievement tests used were classroom tests from an undergraduate course in educational measurement. The tests were each calibrated using both the MAX program (Wright and Panchapakesan, 1969) and the LOGIST program (Wood, Wingersky, and Lord, 1976), yielding for each test 1PL and 3PL item parameter estimates. All the tests had items in common, so item calibration linking was performed using the Least Squares Method (Reckase, 1979) in order to form a large item pool for tailored testing. Then a counterbalanced test-retest design was employed, with each examinee taking both 1PL and 3PL tests in each of two sessions. For both the 1PL and 3PL procedures, items were selected for administration to maximize the value of the information function (Birnbaum, 1968).

The results of this study indicated a number of problem areas in applying tailored testing to multidimensional achievement testing. Both procedures appeared to be inadequate with regard to reliability, with $r = .44$ for the 1PL test and $r = 0.0$ for the 3PL test. In neither case did test information equal the information yielded by the paper-and-pencil test, although the 3PL test came substantially closer than did the 1PL test. Moreover, while the item pool accurately reflected the weighting of the content areas in the paper-and-pencil exam, the items actually selected by the two procedures showed significant deviation from the content distribution of both the item pool and the course exam. It should be noted here that no branching among content areas was attempted. The purpose was to see if selecting items on the basis of information alone would approximate the content area weightings of the item pool.

One other problem that was encountered was nonconvergence of the 3PL maximum likelihood ability estimation in about eight percent of the cases. Recall that it occurred in almost one-third of the cases in the

vocabulary study previously discussed. The substantial reduction in nonconvergence cases was attributed to the use of an item pool of more appropriate difficulty in the achievement testing study.

A number of possible explanations were suggested for the inadequate performance of the 1PL and 3PL procedures. Among these were unstable item parameter estimates due to small sample sizes, a compounding of that instability due to the linking procedures, poor selection of entry points into the item pool, the possibility that latent trait models may not be robust with respect to the violation of the assumption of unidimensionality, and the nonconvergence of the 3PL tailored tests when using maximum likelihood ability estimation.

It is clear from looking at this study that, when applying tailored testing to achievement measurement, careful attention must be paid to the operational characteristics of the procedures. In order to investigate the robustness of the ICC model with respect to violation of the unidimensionality assumption, it is first necessary to eliminate problems such as unstable item calibrations, poor linking procedures, and less than optimal operational characteristics. The present study is an attempt to do just that.

Method

Item Pool Construction

Calibration The test items that were calibrated for use in the item pool were obtained from a series of classroom achievement tests administered in an undergraduate course on educational measurement and evaluation. Items were taken from six different tests of fifty items each, covering the content area of educational evaluation techniques. The tests were calibrated using both the MAX program (Wright and Panchapakesan, 1969), and the LOGIST program (Wood, Wingersky, and Lord, 1976), which yielded the 1PL and 3PL item parameter estimates, respectively. Sample sizes ranged from 148 examinees to 316 examinees. The dates of test administration and sample sizes are presented in Table A-1 of Appendix A.

Linking It would be quite desirable to have a large sample of perhaps 1000 examinees to which a single test of 150 items or more could be administered. This would obviate the need for linking and would provide more stable item parameter estimates. Unfortunately, it is not often possible to administer a test to as many as 1000 examinees at one time. Moreover, for security purposes it is usually necessary to alter a test between administrations, although there may be numerous items in common from one administration of a test to the next. Because of this, it is generally necessary to link together a series of small sample calibrations to get all the item parameter estimates on the same scale. The linking is necessary because the item parameter estimates yielded by the latent trait calibration programs are only invariant to within a linear transformation due to the arbitrary nature of the zero point and the unit of measurement defined by the separate calibrations (Reckase, 1979).

The linking of the 1PL "b" values (item difficulty parameter estimates) was accomplished using the Major Axis Method (Reckase, 1979). Items in common to the tests to be linked were identified, and for each test a mean difficulty value was computed for those items in common. One of the tests was arbitrarily designated as the calibration base, and a second test calibration was linked to it by adding to each item's b-value in the second test a scaling constant equal to the difference between the mean difficulty values that were computed on the common items. The adding of the constant to the second test difficulty values put them on the same scale as the calibration base items. At this point the "b" values for the common items were combined across these two tests using a weighted average procedure based on the sample sizes of the respective calibrations. This same procedure was repeated for all of the remaining tests to be linked using as a calibration base the composite of previously linked tests.

The linking of the 3PL calibrations was done using the Maximum Likelihood Method. This procedure is more fully described by Reckase (1979), and a brief summary here will suffice. This method required the use of the LOGIST program in order to simultaneously calibrate the tests. The test data were first edited into a single large matrix. Items appearing on Test 1 but not on Test 2 were coded as not reached for Test 2, and in this way were not used for the calibration of Test 2. The items in common to the tests ensured that the calibrations were all on the same scale. The full matrix of responses and not reached codes were analyzed to obtain the "a", "b", and "c" parameter estimates.

Item Pool Characteristics The 1PL and 3PL test procedures used identical pools of 183 items. Table 1 summarizes the means, standard deviations, and ranges of the parameter estimates. The correlation between the respective "b" values was $r = .902$. Note that the means and standard deviations of the "b" values for the two calibration procedures are not directly comparable because the origin and unit of measurement set by the two calibration programs are not the same.

The distributions of the item parameter estimates are shown in Figures 1-A, 1-B, 1-C, and 1-D. Probably the most disturbing aspect of these distributions is the positive skewness of the 3PL discrimination values. Approximately 75 percent of the items had discrimination values below .75. Figure 1-B shows that the 3PL difficulty values were also positively skewed. The 3PL item pool did not meet all of the guidelines for item pools as set out by Urry (1977). These guidelines include: item discrimination values should be over .8; item difficulty values should be evenly and widely distributed from about -2.0 to +2.0; guessing values should be less than .3; and there should be at least 100 items in the pool. The 1PL difficulty values (shown in Figure 1-D) were much more uniformly distributed.

Figures 2 and 3 show the information curves for the 1PL and 3PL item pools, respectively. Again, the 3PL curve is positively skewed, with the most information being yielded at the lower range of the ability scale. The 1PL item pool information plots shows a considerably more uniform curve.

Table 1

Descriptive Statistics of Item Parameter Estimates for Tailored Testing Item Pools

	One-Parameter Calibration	Three-Parameter Calibration		
	b_i	a_i	b_i	c_i
Mean	-0.030	.610	-1.674	.180
Median	-0.074	.485	-1.764	.180
S. D.	1.396	.484	3.361	.010
Skewness	-0.284	1.517	1.406	-2.536
Low Value	-5.279	.010	-9.999[a]	.101
High Value	3.052	2.001[b]	14.834	.244

Note. Both pools contained 183 items.

[a]This value was an arbitrary lower limit on the 3PL difficulty parameter estimates.

[b]This value is an upper limit set by the LOGIST program.

Tailored Testing Procedures

The procedures actually used for the tailored testing sessions have been thoroughly described elsewhere (Koch and Reckase, 1978, 1979; Patience, 1977), and so only a brief summary is given here.

Tailored testing procedures have three main components: an item selection routine, an ability estimation technique, and a stopping rule. In this study both the 1PL and the 3PL procedures selected items to maximize the value of the information function (Birnbaum, 1968) at the most recent ability estimate. For the 1PL testing procedure the formula for item information is given by

$$I_i(\theta_j) = \frac{\exp[-(\theta_j - b_i)]}{\{1 + \exp[-(\theta_j - b_i)]\}^2} = \psi(\theta_j - b_i) \qquad (3)$$

where $I_i(\theta_j)$ is the information for Item i at Ability Level θ_j for Examinee j, θ_j and b_i are as previously defined, and $\psi(x)$ is the logistic probability density function. For the 3PL testing procedure the formula for item information is given by

$$I_i(\theta_j) = D^2 a_i^2 \psi[DL_i(\theta_j)] - D^2 a_i P_i(\theta_j) \psi[DL_i(\theta_j) - \log c_i] \qquad (4)$$

Figure 1
ITEM PARAMETER DISTRIBUTIONS

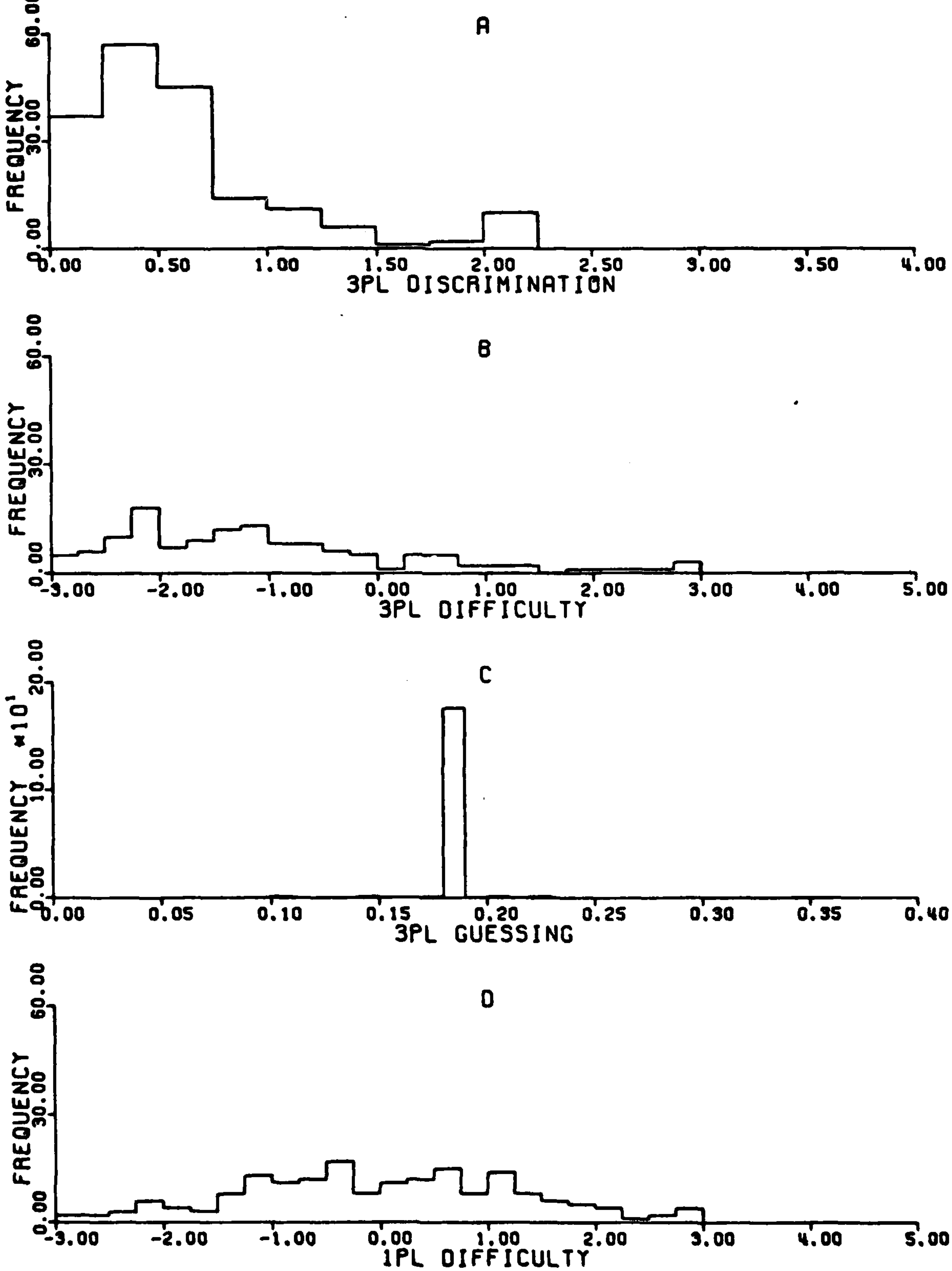
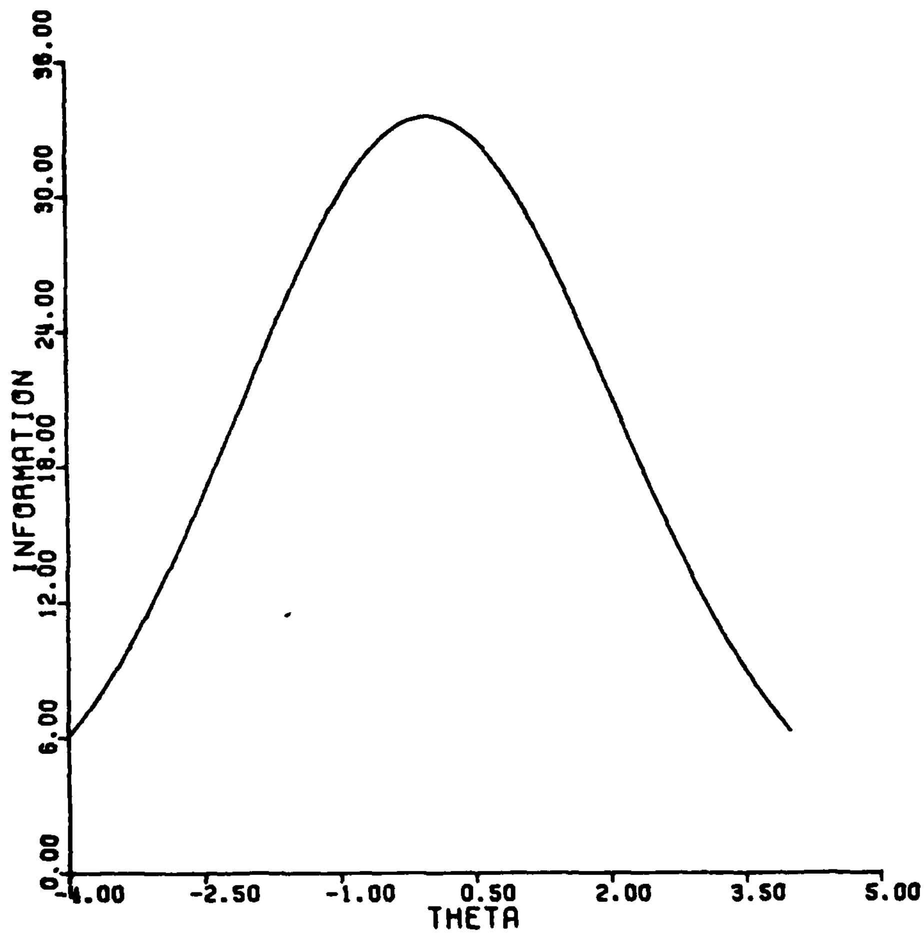


FIGURE 2

INFORMATION CURVE FOR

1PL ITEM POOL



where $I_i(\theta_j)$ is the value of the item information at θ_j, $L_i(\theta_j) = a_i(\theta_j - b_i)$, $P_i(\theta_j)$ is the probability of a correct response to Item i given Ability θ_j, and $\psi(x)$ and the other parameters are as defined earlier. The total test information was defined by Birnbaum (1968) as the sum of the item information values:

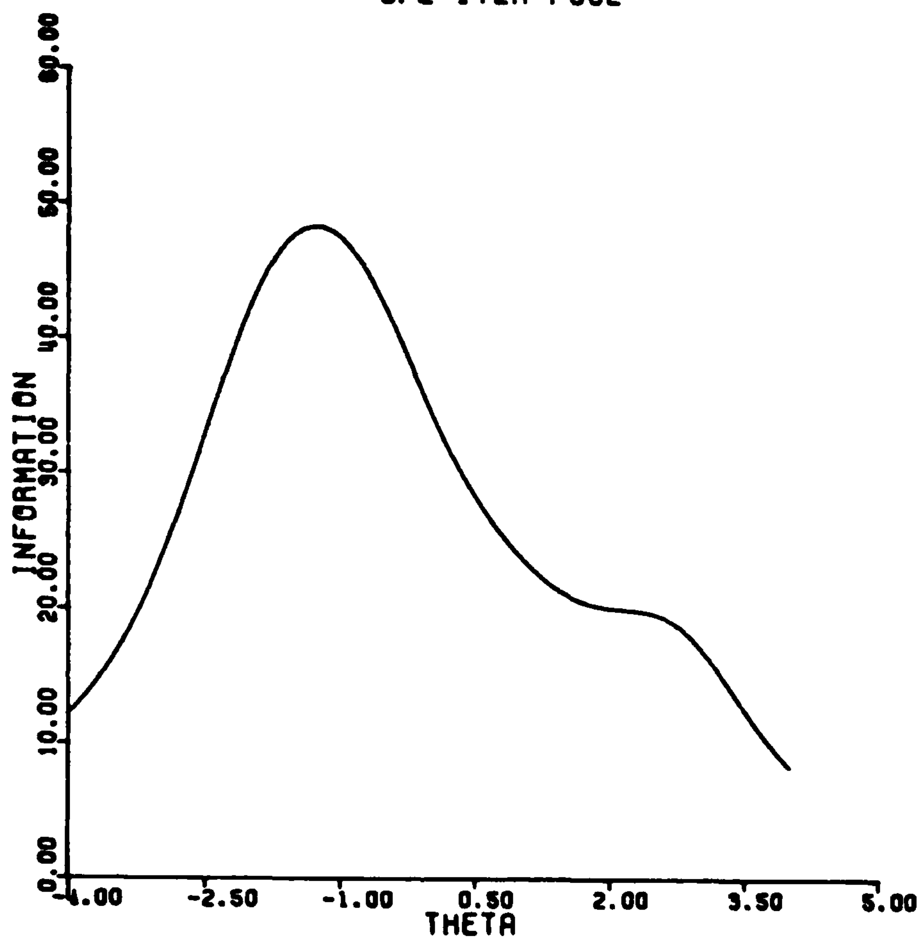
$$I_i(\theta_j) = \sum_{i=1}^{n} I_i(\theta_j) \tag{5}$$

These formulas were used in the tailored testing procedure to compute the information for each item at the examinee's current ability estimate.

FIGURE 3

INFORMATION CURVE FOR

3PL ITEM POOL



The item with the greatest information at that ability estimate was then administered to the examinee, with the provision that the information must be greater than .246 for the 1PL procedure and .65 for the 3PL procedure. These values were chosen based on other research, since they minimize errors in estimation. The information cutoffs were different for the two procedures because the ability scales for the two models are different. If no item were available with information values above these minimums, testing was terminated.

Before testing began no ability estimates were available for the examinees, so initial estimates were assigned to set the starting points in the item pool. The initial ability estimates for this study were set by random assignment to be either -1.856 or -1.500 for the 3PL test, and to be either -.494 or .496 for the 1PL test. These values represent

difficulty values near the medians of the item pool difficulty distributions with one on either side of the median. Two different points were used in order to provide different initial items from one session to the next. The first item was then selected to maximize information at the initial ability estimate. If that item were correctly answered the ability estimate was increased by a fixed stepsize, and if it were incorrectly answered the ability estimate was decreased by a fixed stepsize. This fixed stepsize procedure was used until a maximum likelihood ability estimate, the mode of the likelihood distribution, could be obtained (i.e., when both correct and incorrect responses were obtained). The stepsize used for the 1PL procedure was .693, and for the 3PL procedure it was .4. Each new item was selected to maximize the information at the new ability estimate, with the restriction that no item could be used more than once.

Two stopping rules were used for the testing procedures. The tests were terminated when there were no items left in the item pool with information at the current ability estimate greater than the minimum specified above, or when 20 items had been administered.

Design

This study employed a counterbalanced design using two sessions one week apart. Each session included both a test based on the 1PL model and a test based on the 3PL model. Counterbalancing was achieved by reversal of the order of presentation of the two tests from one session to the next. The test-retest design was used to facilitate reliability comparisons.

During the sessions the tests were administered with no perceptible break between them. The second test was begun immediately after the final ability estimate for the first test was obtained. Since both item pools contained the same items, some of the items in the first test were repeated in the second test. Therefore, examinees were told that they might receive the same item more than once. The tailored tests were administered on Applied Digital Data Systems (ADDS) Consul 980 cathode ray tube terminals connected to an Amdahl 470/V7 computer via time sharing option facilities.

Sample

Examinees were volunteers from an undergraduate introductory measurement course. A total of 88 students participated, 21 male, and 67 female. There were 19 juniors, 67 seniors, and 2 graduate students. The tailored tests were administered shortly after a classroom test over the same content. Examinees were told that the tailored test score would be substituted for the classroom test score if they performed better on the tailored test, and that they would receive extra credit points for completing the requirements of the study.

Attitude Survey

In addition to taking the tailored tests, each examinee was asked to fill out an attitude survey at the end of each session. The survey had 20 items, written in Likert scale format with a five position scale of response alternatives. The surveys were scored with a one for the response least favorable toward the tailored test and a five for the response most favorable.

Analyses

The research questions in this study included a comparison of test-retest reliabilities, goodness of fit, content validity, and total test information functions. In addition, comparisons were made between ability estimates yielded by the 1PL and 3PL procedures, and between the ability estimates and outside criteria. Attitudes of the students toward tailored testing were also determined. Estimated true scores were used in the computation of all the correlations, based on the suggestion of Lord (1979).

The computation of the estimated true scores was accomplished by summing the probabilities of correct responses at the examinee's final ability estimate for all the items in the item pool. The formula for estimated true scores is as follows:

$$\hat{t}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j) \tag{6}$$

where $\hat{t}(\theta_j)$ is the estimated true score for Examinee j.

The reliabilities computed for this study were not strictly test-retest reliabilities, but rather a mixture of test-retest and equivalent forms reliabilities since the tests in one session were not identical to tests taken in the other session. The reliabilities were compared using a t-test based on Fisher's r to z transformation.

The total test information analyses were done to compare the relative efficiencies (Birnbaum, 1968) of the tailored testing procedures with respect to the course exam. The relative efficiency was the ratio of the information provided by the tailored test at a particular ability to the information of the traditional paper-and-pencil course exam at the same ability. Plots were drawn of the relative efficiency curves for the two tailored testing models based on sample cases selected from across the entire range of the tailored testing ability estimates.

Other analyses run on the data included a series of correlational analyses. For instance, correlations between the 1PL and 3PL ability estimates were run using estimated true scores, as were correlations between the ability estimates and course exam scores. The exams that were correlated with estimated true scores included the course exam over the same content area as the tailored tests as well as two other course exams and the sum of all the course exams. The objective of all these correlational

analyses was to see whether the 1PL and 3PL tests measured the same thing, and whether one test correlated more highly with the outside criteria. The correlations of the tailored test scores and the outside criteria were an indication of concurrent validity. In addition to the above analyses, descriptive statistics were compiled, including the average test lengths, the average test difficulties, and the number of items used from each item pool, for both sessions of the 1PL and 3PL tests.

The goodness of fit statistic used in this study was the mean square deviation, calculated by summing over examinees the squared differences between the actual responses to the items and the expected responses to the items (probability of a correct response) as predicted by the models. The formula for the MSD statistic is

$$MSD_j = \sum_{i=1}^{n} \frac{(u_{ij} - P_i(\theta_j))^2}{n_j} \tag{7}$$

where MSD_j is the mean squared deviation for Examinee j, u_{ij} is the actual response to Item i by Examinee j, $P_i(\theta_j)$ is the probability of a correct response to Item i by Examinee j determined from the model using the final ability estimate and the estimated item parameters, and n_j is the number of items in the tailored test for Examinee j. The MSD statistic was computed for a systematic sample of 29 examinees from across the ability range. The 1PL and 3PL tests were compared using the MSD statistic as the dependent variable in a dependent t-test.

Content validity analyses were done to determine the degree to which the item pools and the tailored tests accurately represented the content breakdown of the traditional test. Actual and expected frequencies of content samplings were compared using a χ^2 statistic. Since the argument was presented that achievement tests are typically multidimensional, factor analyses were also run on the course exam to determine the factor structure of the test. Principal components analyses with varimax rotations were employed.

Student attitudes were analyzed using data from the surveys administered at the end of each session. The first analysis run on the response data was a principal components factor analysis followed by a varimax rotation. Once the factor structure was determined, attempts were made to label factors and compare them with the factors from previous administrations of the scale reported by Koch and Reckase (1978, 1979). Coefficient alpha reliabilities were calculated for each factor as well as for the total scale. Response frequencies for the five scale positions were tabulated for both sessions to summarize student attitudes toward tailored testing. Also, multivariate analyses were run to determine if there were significant change in attitudes from one session to the next.

Results

Reliability

Table 2 contains the correlation matrix obtained from intercorrelating the ability estimates yielded by the two models used in the tailored testing sessions. The correlation of r = .57 between the ability estimates from the first 1PL test (1PL 1) and the ability estimates from the second 1PL test (1PL 2) was the reliability for the 1PL procedure. The reliability for the 3PL procedure, r = .62, was higher, but not significantly so. The KR-20 reliability of the traditional paper-and-pencil course exam was r = .60. The reliabilities of the tailored tests were actually substantially higher than the reliability of the conventional test, since normally a KR-20 reliability would be expected to be higher than a test-retest reliability. Also, it should be noted that the tailored tests were less than half as long as the conventional test.

Table 2

Ability Estimate Correlations

Model	Session	1	2	3	4
1. 1PL	1	1.00	.57	.35	.42
2. 1PL	2		1.00	.38	.44
3. 3PL	1			1.00	.62
4. 3PL	2				1.00

Table 3 shows that the tailored test reliabilities were even higher when estimates true scores were used in place of ability estimates. Using estimated true scores, the 1PL reliability was r = .62 and the 3PL reliability was r = .71.

Table 3

Ability Estimate Correlations Using Estimated True Scores

Model	Session	1	2	3	4
1. 1PL	1	1.00	.62	.36	.44
2. 1PL	2		1.00	.41	.52
3. 3PL	1			1.00	.71
4. 3PL	2				1.00

Information

The relative efficiency comparison of the total test information for the 1PL and 3PL procedures is shown in Figure 4. The horizontal

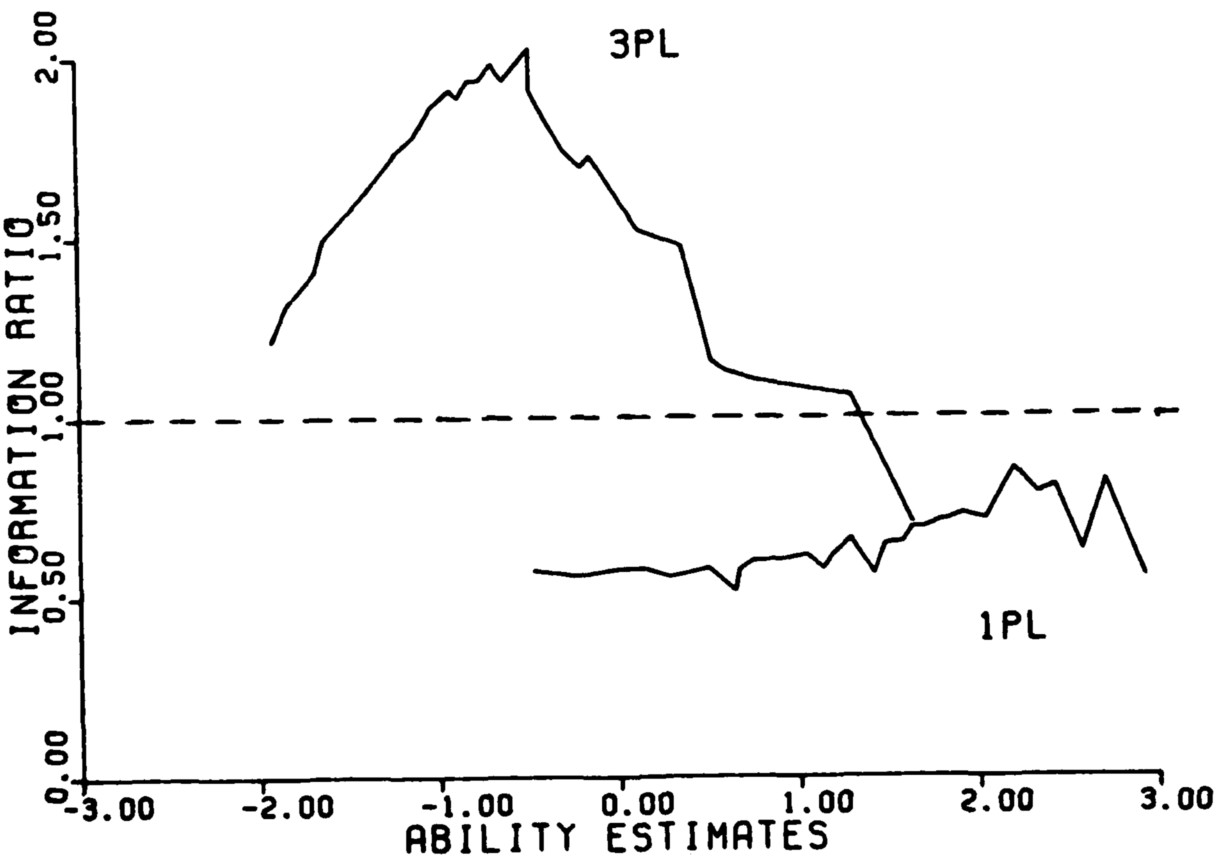
broken line represents the relative efficiency of the course exam, which was used as a standard for comparing the two procedures. It should be noted that the ability scale for the 1PL model is not the same as the ability scale for the 3PL model. Thus the plots are not comparable on a point by point basis. However, an overall visual examination of the plots of information curves for the two models is still possible.

FIGURE 4

TOTAL TEST INFORMATION

COMPARISON OF 1PL AND

3PL TAILORED TESTS



Perhaps the most significant result of this comparison is that the 3PL procedure not only yielded more information than the 1PL procedure, but in the ability estimate range of -2.0 to +1.5 the 3PL procedure also yielded more information than the 50 item paper-and-pencil test. It is important to point out that the 3PL procedure performed best in that range of ability estimates where most of the examinees were classified, while the 1PL procedure had its highest relative efficiency at the upper end of the range of ability estimates, where few examinees were classified. See Appendix B for the distribution of ability estimates for both the 1PL and 3PL procedures.

Goodness of fit

Table 4 presents the results of the goodness of fit comparison of the 1PL and 3PL models using the MSD statistic. MSD values were computed for 29 cases for each model, along with means, standard deviations, and the results of a dependent t-test analysis of the data. The results of the t-test indicated that the 3PL model fit the observed responses significantly better than the 1PL model ($p < .001$).

Correlational Analyses

Table 5 and 6 show the correlations of the traditional course exam scores and total course scores (the sum of the course exam scores) with the tailored testing ability estimates and with the estimated true scores, respectively. The differences between the correlations of the 1PL and 3PL ability estimates with Exam I were not significant, while the 1PL correlation was significantly higher than the 3PL correlation with respect to the total score for the first session ($p < .05$) but not for the second. The correlations did not change significantly when estimated true scores were used instead of ability estimates.

One interesting result that is shown in Table 5 is that the 1PL 1 ability estimates correlated significantly higher with Exam II than with Exam I ($p < .05$). Moreover, both the 1PL 1 and the 1PL 2 ability estimates correlated higher with the total course score than with Exam I ($p < .01$ for 1PL 1, $p < .05$ for 1PL 2). Remember that Exam I was the course exam over the same material as the tailored tests. One possible explanation for this is that the KR-20 reliabilities of Exam II and the total course score were higher than the reliability of Exam I. The reliability of the total course score was computed according to a method suggested by Lord and Novick (1968, pp. 203-204). These reliabilities are shown in Table 5.

Descriptive Statistics

Table 7 presents descriptive statistics for both sessions of the 1PL and the 3PL tailored tests. The mean number of items administered indicates that the 1PL tests tended to be longer than the 3PL tests, and that many of the 1PL tests went the maximum of 20 items. The mean pro-

Table 4

Goodness of Fit Comparison
Using the MSD Statistic

Observations	One-Parameter MSD	Three-Parameter MSD
1	.1887	.1103
2	.1833	.0142
3	.1863	.0832
4	.2085	.1894
5	.2123	.1226
6	.2087	.1394
7	.1853	.0349
8	.2107	.1137
9	.2133	.2273
10	.2174	.1216
11	.1923	.2405
12	.2219	.2515
13	.2120	.1826
14	.2197	.2171
15	.2192	.0728
16	.2033	.1712
17	.2176	.1984
18	.2124	.2024
19	.2122	.2305
20	.2015	.1616
21	.2095	.0457
22	.1883	.1309
23	.2230	.2107
24	.1367	.0235
25	.2086	.1751
26	.2177	.2281
27	.2087	.1330
28	.2137	.0994
29	.2097	.1693
$\bar{x}$	.2049	.1483
$S_{\bar{x}}$	.0425	.0740

$\underline{t}_{(28)} = 5.082 \qquad (\underline{p} < .001)$

portion of items answered correctly shows that the 3PL procedure administered items that were, overall, easier than those items administered by the 1PL procedure.

An important effect related to the 3PL item discrimination parameter estimates was that only 25 items from the 183 items in the 3PL item pool were used by the 3PL testing procedure. On the other hand the 1PL procedure used 120 items from the 183 items in the 1PL item pool. Figure

Table 5

Correlations of Ability Estimates with Traditional Course Exams

Traditional Course Exam	KR-20 Reliability	Tailored Testing Model and Session 1PL 1	1PL 2	3PL 1	3PL 2
Exam I*	.60	.42	.49	.39	.42
Exam II	.76	.58	.46	.36	.47
Exam II	.64	.36	.35	.38	.44
Total Score	.75	.68	.63	.45	.52

*Exam I was over the same content area as the tailored tests.

Table 6

Correlations of Estimated True Scores with Traditional Course Exams

Traditional Course Exam	Tailored Testing Model and Session 1PL 1	1PL 2	3PL 1	3PL 2
Exam I*	.42	.49	.40	.42
Exam II	.58	.46	.36	.44
Exam III	.37	.33	.40	.44
Total Score	.68	.62	.46	.51

*Exam I was over the same content area as the tailored tests.

Table 7

Tailored Test Descriptive Statistics

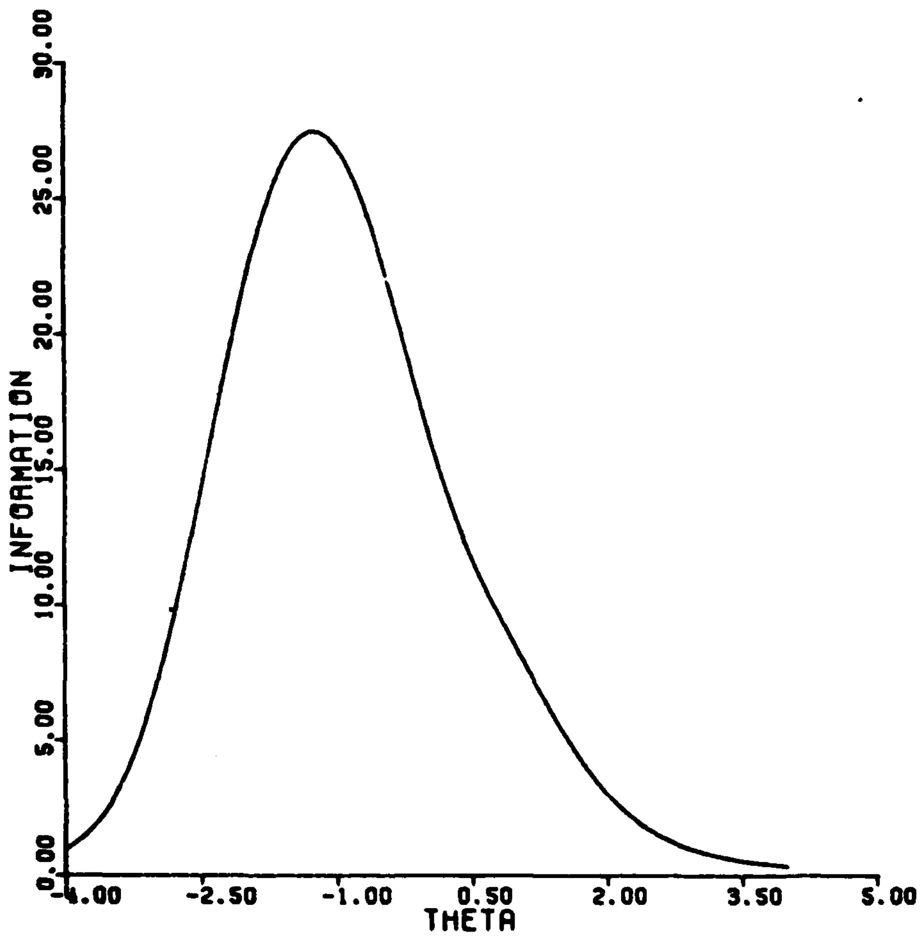
Variable	One-Parameter Tailored Test		Three-Parameter Tailored Test	
	Session 1	Session 2	Session 1	Session 2
Mean # of items administered	19.09	18.11	16.23	15.32
Mean # of items correct	11.07	10.30	12.15	11.71
Mean proportion of items correct	.58	.57	.75	.76
Mean of ability estimates	1.37	1.50	-.53	-.36
S.D. of ability estimates	.67	.92	.74	.83

5 shows the information curve for the 25 items that were used from the 3PL item pool. The plot shows that the most information yielded by this reduced pool was at the lower range of abilities. In fact, for ability estimates over +2.0 there were virtually no items available with information above the information cutoff.

FIGURE 5

INFORMATION CURVE FOR 25

ITEMS USED FROM 3PL POOL



Content Validity

Table 8 shows the results of the content validity analysis for both tailored testing models. The Chi-Square test indicated that both the 1PL and the 3PL item pools accurately reflected the weighting of the content areas specified in the table of specifications for the paper-and-pencil course exam (see Table A-2 in Appendix A). However, the number of items administered by content area for a systematic sample of 21 1PL tailored

tests and 20 3PL tailored tests showed significant lack of fit to both the item pools and the course exam. Also, the content distributions of the 1PL and 3PL tailored test items were significantly different. It should be noted that no attempt was made in the tailored testing procedures to branch among the content areas. The object was to see if selecting items for administration on the basis of information alone would approximate the content area weightings of the item pools and the course exam.

Attitude Survey

Attitude Scale Characteristics Table 9 shows the varimax rotated factor loading matrix obtained from a principal components analysis of the first administration of the attitude scale. There were six factors present with eigenvalues greater than one, accounting for 62.5 percent of the variance. A subjective examination of the items loading on each factor resulted in the following factor labels:

factor I - cathode ray tube (CRT) characteristics
factor II - perceived test performance/test satisfaction
factor III - motivation
factor IV - anxiety
factor V - test pace
factor VI - time pressure/item easiness

The items appearing on the attitude scale are listed in Appendix C.

Table 10 shows the rotated factor loading matrix obtained from the analysis of the second administration of the attitude scale. This time there were five factors present with eigenvalues greater than one, accounting for 62 percent of the variance. After a subjective examination, these factors were given the following labels:

factor I - perceived test performance/test satisfaction
factor II - motivation
factor III - anxiety/time pressure
factor IV - miscellaneous
factor V - CRT characteristics/item easiness

Factor analysis results obtained from the two attitude scale administrations differed somewhat. For instance, in the first administration of the scale, anxiety and time pressure items loaded on separate factors, while in the second administration they formed a single factor. Another difference was that in the first administration, item easiness items loaded with time pressure items, while in the second administration item easiness items loaded with CRT characteristics items. Also, in the first administration Item 11 loaded by itself, while in the second administration it was joined by three other items in a factor of assorted item types, labelled here as miscellaneous.

A multivariate analysis of variance (MANOVA) was performed to determine whether the mean scores on each item were different over the two administrations of the attitude scale. The results of the MANOVA indicated that there were no significant changes. This implied that, regardless of the changes in factor structure, student attitudes toward tailored testing did not change from one administration to the next.

Table 8

Test Items by Content Area for the Course Exam, Item Pools, and Tailored Tests

	Course Exam Items		Items in 1PL Pool		Items in 3PL Pool		Items in 21 1PL Tailored Tests		Items in 20 3PL Tailored Tests	
	Number	%	Number	%	Number	%	Number	%	Number	%
Anecdotal Records	5	10.0	18	9.8	18	9.8	62	15.3	0	0
Behavioral Objectives	5	10.0	20	10.9	20	10.9	43	10.6	47	15.5
Checklists	5	10.0	18	9.8	18	9.8	37	9.1	21	6.9
Peer Appraisals	2	4.0	5	2.7	5	2.7	6	1.5	14	4.6
Planning Tests	3	6.0	12	6.6	12	6.6	38	9.4	18	5.9
Rankings	3	6.0	9	4.9	9	4.9	4	1.0	20	6.6
Ratings	6	12.0	25	13.7	25	13.7	76	18.7	64	21.1
Selection Items	8	16.0	30	16.4	30	16.4	55	13.6	59	19.5
Self Report	2	4.0	8	4.4	8	4.4	12	3.0	14	4.6
Supply Items	5	10.0	20	10.9	20	10.9	32	7.9	9	3.0
Table of Specifications	6	12.0	18	9.8	18	9.8	41	10.1	37	12.2
Total	50		183		183		406		303	

Note. Below are the Chi-Square values for several comparisons. The critical value for rejection of adequate fit is $\chi^2_{(10)} > 18.31$ at $\alpha = .05$.

1. Course exam items vs. items in 1PL pool, $\chi^2 = 4.431$
2. Course exam items vs. 1PL tailored test items, $\chi^2 = 55.078$
3. 1PL pool items vs. 1PL tailored test items, $\chi^2 = 43.139$
4. Course exam items vs. items in 3PL pool, $\chi^2 = 4.431$
5. Course exam items vs. 3PL tailored test items, $\chi^2 = 80.878$
6. 3PL pool items vs. 3PL tailored test items, $\chi^2 = 77.662$
7. 1PL tailored test items vs. 3PL tailored test items, $\chi^2 = 89.02$

Table 9

Principal Components Analysis
Varimax Rotated Factor Pattern for
First Attitude Survey Administration

	Factor					
Item No.	I	II	III	IV	V	VI
1	-.06	.24	-.28	.47	.45	.03
2	.15	.09	.06	.01	-.13	.76
3	-.23	.68	-.12	.01	-.40	.02
4	.23	-.09	.19	.52	-.10	.43
5	-.08	.11	.78	.06	-.06	.23
6	-.19	.64	.33	.29	.24	.03
7	.22	.70	.04	.12	-.14	.12
8	.71	.01	.08	-.00	-.09	.05
9	.14	-.10	.26	-.20	.53	.58
10	.02	.14	-.13	.72	-.03	-.18
11	.31	-.04	.13	.07	-.60	.19
12	.25	.64	-.10	-.03	.32	-.13
13	.74	.12	.06	.19	.10	.04
14	.11	.65	.15	.16	.22	-.31
15	.41	.20	-.09	.62	-.05	.24
16	.72	.17	.13	-.06	-.20	.23
17	.20	.78	.30	-.02	-.05	.27
18	-.31	-.05	.43	.49	-.01	-.10
19	.18	.04	.70	-.11	-.16	.02
20	.32	.19	.69	-.09	.16	-.00

Note. The underlined values indicate the highest loadings of an item on a factor. Broken underlines indicate other high loadings.

A comparison of the results of the attitude scale administrations for this study with results from previous administrations of the scale indicated several differences. For instance, in the earliest administration of the scale (Koch and Reckase, 1978) anxiety and time pressure items loaded on separate factors, while in a subsequent study (Koch and Reckase, 1979) they formed a single factor. In the present study, they loaded on separate factors in the first administration, and on the same factor in the second administration. In both of the earlier studies perceived test performance and test satisfaction items loaded on separate factors, while in the present study they formed a single factor in both administrations.

Two types of reliability measures were computed for the attitude scale. First, a test-retest reliability coefficient was computed between the sets of total attitude scores for the two administrations. A value of $r = .71$ was obtained for this reliability measure. The second type of reliability measure calculated for the attitude scale was a coefficient alpha reliability. Coefficient alpha reliabilities were computed for each

Table 10

Principal Components Analysis
Varimax Rotated Factor Pattern for
Second Attitude Survey Administration

Item No.	Factor I	II	III	IV	V
1	.20	-.40	.14	.57	.05
2	.03	.25	.66	-.22	-.04
3	.23	.19	.01	.72	.19
4	-.17	.39	.64	.17	.25
5	.25	.77	-.11	-.08	.08
6	.58	-.11	.15	.43	-.25
7	.79	.25	.07	.08	-.03
8	-.10	.15	.18	.13	.79
9	-.03	.34	-.12	-.26	.45
10	.23	-.27	.75	.27	-.02
11	.26	-.05	.20	-.65	.24
12	.65	.03	-.10	.04	.40
13	.07	.00	.72	.07	.45
14	.51	.02	.21	.69	-.09
15	.36	-.39	.58	-.03	.10
16	.27	.00	.33	-.14	.64
17	.83	.20	.17	.11	.03
18	.09	.59	-.03	.17	-.12
19	.22	.60	.12	-.11	.24
20	-.04	.64	.16	.01	.24

Note. The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

factor and for the total scale for both administrations of the instrument. The results are shown in Table 11 for the first administration and in Table 12 for the second administration. Overall these reliabilities were fairly high. However, for the first administration, the reliability of the time pressure/item easiness factor was relatively low. Note that in the second administration these two item types did not load together. In the second administration the only factor not having a high reliability coefficient was the miscellaneous factor.

Item discrimination indices were calculated for the items on the attitude survey by correlating individual item scores with the total scores for each examinee. These values are shown in Table 13. Discriminations were relatively constant across the two administrations, with the exception of Item 10.

Attitude Scale Results Responses obtained from the administration of the attitude scale are summarized in Table 14. Response percentages for the five categories for each item are shown for both administrations.

Table 11

Coefficient Alpha Reliabilities for Attitude Survey Factors and Total Scale for Session I

	Factor Labels	Items	Coeff. α
I.	CRT Characteristics	8, 13, 16	.69
II.	Perceived Test Performance/ Test Satisfaction	3, 6, 7, 12, 14, 17	.79
III.	Motivation	5, 19, 20	.66
IV.	Anxiety	1, 4, 10, 15, 18	.52
V.	Time Pressure/Item Easiness	2, 9	.28
	Total Scale	all 20 items	.75

Note. Item 11 loaded on its own factor, so no coefficient α could be calculated for it alone.

Table 12

Coefficient Alpha Reliabilities for Attitude Survey Factors and Total Scale for Session II

	Factor Labels	Items	Coeff. α
I.	Perceived Test Performance/ Test Satisfaction	6, 7, 12, 17	.77
II.	Motivation	5, 18, 19, 20	.66
III.	Anxiety/Time Pressure	2, 4, 10, 13, 15	.74
IV.	Miscellaneous	1, 3, 11, 14	.22
V.	CRT Characteristics/Item Easiness	8, 9, 16	.55
	Total Scale	all 20 items	.77

Overall the results of the attitude survey were positive regarding attitudes toward the tailored testing situation. Examinees indicated that they felt less time pressure when taking the tailored test than when taking the conventional test. However, responses indicated a split over whether the examinees felt that they did well on the tailored test, and many examinees remained neutral on those items dealing with test performance. Examinees indicated that they were motivated to do well on the test, but felt little anxiety or stress. The examinees responded that they felt comfortable with the CRTs, and that the screens were not difficult to read. Test items were apparently perceived as neither too difficult nor too easy, but examinees were split over whether they believed the tailored tests reflected their true knowledge of the material. No significant correlations were found between the attitude scores and the ability estimates.

Table 13

Discrimination Indices for Attitude Scale
Items for Two Test Sessions

Item No.	Session I	Session II
1	.26	.28
2	.41	.41
3	.29	.43
4	.45	.52
5	.41	.36
6	.48	.35
7	.59	.57
8	.46	.45
9	.18	.18
10	.28	.52
11	.31	.28
12	.41	.51
13	.54	.65
14	.44	.51
15	.59	.46
16	.56	.58
17	.72	.65
18	.22	.28
19	.33	.46
20	.46	.37

Discussion

In order to fully understand the results of the research reported here, the results of three tailored testing studies should be kept in mind: (a) the application of tailored testing models to a vocabulary test (Koch and Reckase, 1978), (b) a previous attempt to apply tailored testing models to achievement testing (Koch and Reckase, 1979), and (c) the current study. The first study, using the vocabulary test, was successful, but the success was not surprising, since the vocabulary test used was highly unidimensional. However, nonconvergence of the ability estimates was found to be a problem. The high nonconvergence rate was felt to be due to the inappropriate difficulty of the item pool. When an attempt was made to apply tailored testing to a multidimensional achievement test, the nonconvergence problem was reduced through the use of items of appropriate difficulty, but other problems were encountered (e.g., low reliabilities), and the attempt at application was unsuccessful.

There were indications that the lack of success in this first achievement testing study might have been due to factors other than the multidimensional nature of the test, such as the linking procedures used with the calibrations. The current study, in which improvements were made in the operational characteristics of the tailored testing procedures

Table 14

Attitude Scale Response Percentages for Item Alternatives over Both Sessions

Item No.	Session 1					Session 2				
	SA	A	N	D	SD	SA	A	N	D	SD
1	6	43	16	27	8	5	20	17	39	19
2	32	45	9	10	3	20	49	15	15	1
3	5	53	34	6	2	7	60	25	8	0
4	1	9	6	47	38	1	6	7	49	38
5	27	48	17	8	0	18	53	15	11	2
6	1	26	63	10	0	0	39	55	7	0
7	5	27	26	39	3	1	39	30	31	0
8	13	23	13	35	17	6	24	10	42	18
9	6	72	23	0	0	6	73	22	0	0
10	0	14	8	44	34	3	6	6	47	39
11	17	53	8	18	3	13	51	10	19	7
12	8	48	24	20	0	2	40	31	26	1
13	3	5	3	58	31	1	6	6	52	35
14	0	15	48	38	0	1	17	50	32	0
15	38	43	11	7	1	31	58	3	6	2
16	25	55	9	10	1	23	55	11	10	1
17	1	38	35	22	5	0	28	40	27	5
18	1	38	19	31	11	2	35	22	34	7
19	0	0	5	60	35	0	2	5	66	27
20	1	1	5	51	42	0	2	9	56	33

Note. SA = Strongly Agree, A = Agree, N = Neutral, D = Disagree, SD = Strongly Disagree. For a list of the actual items, see Appendix C.

and the linking procedures, demonstrated that tailored testing could be successfully applied to a multidimensional test, if reliability and information functions were used as criteria. Indeed, the current study employed virtually the same item pool as the first tailored achievement testing study, but the results were quite different. The difference between these two achievement studies was not in the dimensionality of the item pool, but in the operational characteristics of the procedures employed. The changes that were made and their effects will now be discussed.

Reliability

A number of changes implemented during the design of the current study probably contributed to the gain in the tailored test reliabilities over the previous tailored testing achievement study. One such change was the improvement of the linking procedures that were employed. The

1PL item parameter estimates were linked using the same method as was used in the previous studies. However, previously the linking had been done by hand, while this time computer programs were used to perform the linking. Therefore, any computational errors that might have occurred in linking should have been eliminated. For this study the 3PL calibrations were linked using the Maximum Likelihood Method, rather than the Least Squares Method that had been used earlier. Again linking was performed by computer programs instead of by hand. These improvements in linking provided more accurate item parameter estimates for the items in the pools.

Another important change was that larger sample sizes were used for item calibration. Sample sizes used ranged from 148 to 314, with a mean sample size of 226.5. These were not much larger than the sample sizes used in previous studies for the 1PL calibrations, but they were somewhat larger than the sample sizes used previously for the 3PL calibrations. In the previous tailored achievement testing study the 1PL sample sizes ranged from 96 to 314, with a mean of 212.82, while 3PL sample sizes ranged from 97 to 314, with a mean sample size of 195.4. The larger sample sizes may have yielded more stable parameter estimates than the previous smaller sample sizes, although Reckase (1977) found that these sample sizes were still inadequate for the 3PL calibration.

Other important changes were in the procedures used in administering the tailored tests. For instance, entry points (initial ability estimates) for the 3PL procedures were set at the difficulty values on either side of the median of the item pool difficulty distribution. In earlier studies the entry points were arbitrarily set to be $\pm.5$, because the item pool was assumed to be centered around zero. This was found to not be the case. By using entry points near the median of the difficulty distribution more items were available within the fixed stepsize in either direction. Also, the fixed stepsize that was used was .4, rather than the .693 that had previously been used for the 3PL procedure. This helped to avoid the previously encountered problem of moving through the item pool too quickly, resulting in premature termination of the test. These changes in the entry points and fixed stepsize for the 3PL procedure were important factors in the virtual elimination of the problem of nonconvergence and, together with the improved calibrations and linkings, probably accounted for the higher reliabilities of the tailored tests.

Information

In looking at the information yielded by the tailored tests it should be remembered that the tailored tests were less than half the length of the classroom test. Since total test information was the sum of the individual item information, a drop in total information would be expected when considering a shorter test. Despite this, the 1PL tailored test yielded almost as much information as the classroom test, and the 3PL tailored test yielded more information than the classroom test over most of the ability range.

Goodness of fit

The superior fit of the 3PL model indicated that the 3PL tailored tests demonstrated better 'person' fit than did the 1PL tests. It was no surprise that the three-parameter model fit observed response data better than the one-parameter model. A model with three parameters has more flexibility in fitting data than a model with only one parameter. Such a finding is consistent with the findings of previous studies (Koch and Reckase, 1978, 1979).

Correlational Analyses

In correlating the tailored testing ability estimates with the outside criterion variables, it was found that the 1PL 1 ability estimates correlated significantly higher with Exam II than with Exam I. Also, both the 1PL 1 and 1PL 2 ability estimates correlated significantly higher with the total course score than with Exam I. This is somewhat surprising, since Exam I was the course exam over the same content as the tailored tests. However, this might be explained by examining the reliabilities of the course exams. The KR-20 reliabilities of Exam II and the total course score were higher than the KR-20 reliability of Exam I. The lower reliability of Exam I might be limiting the magnitude of the correlations that can be obtained using that test. Of course, this would be true for correlations of Exam I with both the 1PL and 3PL ability estimates. One reason why this effect appeared with the 1PL ability estimates and not the 3PL ability estimates might be that since the 1PL calibrations are based on the sum of the factors the 1PL tests might have had factors in common with Exam II. The 3PL calibrations are based on the dominant factor, which the 3PL tests would have in common with Exam I but not Exam II. Any sharing of factors between the 1PL tests and Exam II would have caused that correlation to be higher than the correlation between the 3PL ability estimates and Exam II. However, these explanations are only conjecture, and further studies are needed to determine if these anomalous results can be replicated.

Content Validity

The content validity results clearly indicated that, even though the item pools reflected content area weightings proportionate to the classroom test, the tailored test item selection procedures did not maintain these content weightings. For the 3PL procedure this was not surprising. High item discriminations were not distributed evenly across content areas and, since the 3PL procedure selected items on information, those content areas having no highly discriminating items were not represented at all. Content areas with several high discriminators were weighted too heavily relative to the table of specifications. The reason for this imbalance in the distribution of item discriminations was probably caused by the loading of the highly discriminating items on the dominant factor. Previous research (Reckase, 1977) had indicated that the 3PL model calibrates items based on the dominant factor in the test, resulting in low discrimination values for items loading on the remaining factors, while the 1PL

procedure calibrates items based on the sum of the factors. Given these contrasting tendencies, it is not surprising that the 3PL tailored tests used only 25 items out of 183, whereas the 1PL tailored tests used 120 items out of 183. This effect is reflected in the low correlations between the 1PL and 3PL ability estimates shown in Table 2.

For the 1PL procedure, however, item discriminations were assumed to be equal, so the result was somewhat surprising. A possible explanation is that content areas are not uniformly distributed across the difficulty scale. The results indicated that, if content areas were to be weighted appropriately, some type of intercontent area branching scheme would have to be employed. An alternative to branching might be to administer tailored tests over unidimensional subtests and to report a profile of scores. Of course, this alternative carries with it the problem of identifying unidimensional subtests, as well as determination of a total score when one is desired.

Attitude Survey

The attitude scale results were generally favorable toward tailored testing. However, there was no evidence to indicate any interaction between either student motivation or anxiety levels and student test performance. These findings were consistent with the findings of the previous study, which found no significant correlation between attitudes of the students toward the tailored tests and their performance. It should be emphasized that these studies were performed using college juniors and seniors, most of whom were females, and the results may not generalize to other groups.

The factor structure of the attitude scale appeared to be unstable. Not only did a number of items switch factors, but the factors themselves changed both in number and in their nature. For instance, a number of items that loaded on separate factors in the first administration of the scale grouped together in the second administration to form a new factor that did not occur in the first administration. The items that loaded on this new factor, labelled miscellaneous, were items that did not appear to be related at all. One possible reason for the unstable factor structure of the scale was the small sample size. For a scale of 20 items, 88 is not an adequate number of subjects to obtain a stable structure. It is interesting to note that when an analysis of the factor structure of the attitude scale using the skree technique was performed the results were ambiguous. The plot of eigenvalues by the factors is shown in Appendix D. The number of factors determined using the eigenvalue-greater-than-one rule gave probably as good an indication of the number of factors as that obtained from the skree plot.

Summary and Conclusions

Past studies indicated that there might be serious problems with the application of tailored testing to multidimensional achievement test-

ing. However, there was some evidence that those findings were the result of poor item calibration, linking procedures, and test administration procedures. The present study showed that if sufficient attention was paid to establishing proper operational characteristics, tailored testing could be successfully applied to multidimensional achievement tests to the extent that they yielded high reliabilities and information.

The results of this study indicate that tailored test reliabilities for both the 1PL and 3PL procedures were probably higher than the reliability of the classroom test. The information yielded by the 1PL test was almost as high as the classroom test information, and the 3PL test information was higher than either one. The fit of the two models to the response data showed that the 3PL model fit the data better than the 1PL model. Neither procedure, however, had adequate content validity. In summary, these results showed that tailored testing is a viable procedure for achievement testing, with the exception of content validity, and that the 3PL model appears to be the model of choice.

REFERENCES

Bejar, I. I., Weiss, D. J. and Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Brown, J. M. and Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Koch, W. R. and Reckase, M. D. A live tailored testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). Columbia: University of Missouri, Department of Educational Psychology, 1978.

Koch, W. R. and Reckase, M. D. Problems in application of latent trait models to tailored testing (Research Report 79-1). Columbia: University of Missouri, Department of Educational Psychology, 1979.

Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Lord, F. M. Personal communication, June, 1979.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Patience, W. M. Description of components in tailored testing. Behavior Research Methods and Instrumentation, 1977, 9, 153-157.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: a comparative study (Research Report 77-1). Columbia: University of Missouri, Department of Educational Psychology, 1977.

Reckase, M. D. Item pool construction for use with latent trait models. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Paper presented at the Annual Meeting of the American Psychological Association, Chicago, August, 1975.

Urry, V. W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS Research Memorandum RM-76-6). Princeton, New Jersey: Educational Testing Service, June, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

APPENDIX A

Table A-1

Administration Dates and Sample Sizes of Achievement Tests Calibrated for Tailored Testing Usage

Date	Sample Size
9-76	177
2-77, 4-77	314
9-77, 10-77	202
2-78, 4-78	309
9-78, 11-78	209
2-79	148

Note. Dates given in month and year.

Table A-2

Table of Specifications for Exam I

Content Areas	Knowledge of Terms and Techniques	Application of Techniques	Analysis, Synthesis, and Evaluation of Techniques	Totals
Planning the Test	1	1	1	3
Behavioral Objectives	1	2	2	5
Table of Specifications	2	2	2	6
Anecdotal Records	1	2	2	5
Rating Scales	2	2	2	6
Checklists	1	2	2	5
Rankings	1	1	1	3
Peer Appraisals	1	1		2
Self Reports	1	1		2
Selection Items	2	3	3	8
Supply Items	1	2	2	5
Totals	14	19	17	50

Appendix B

ABILITY ESTIMATE FREQUENCY

DISTRIBUTIONS

FIGURE B-1

IPL 1

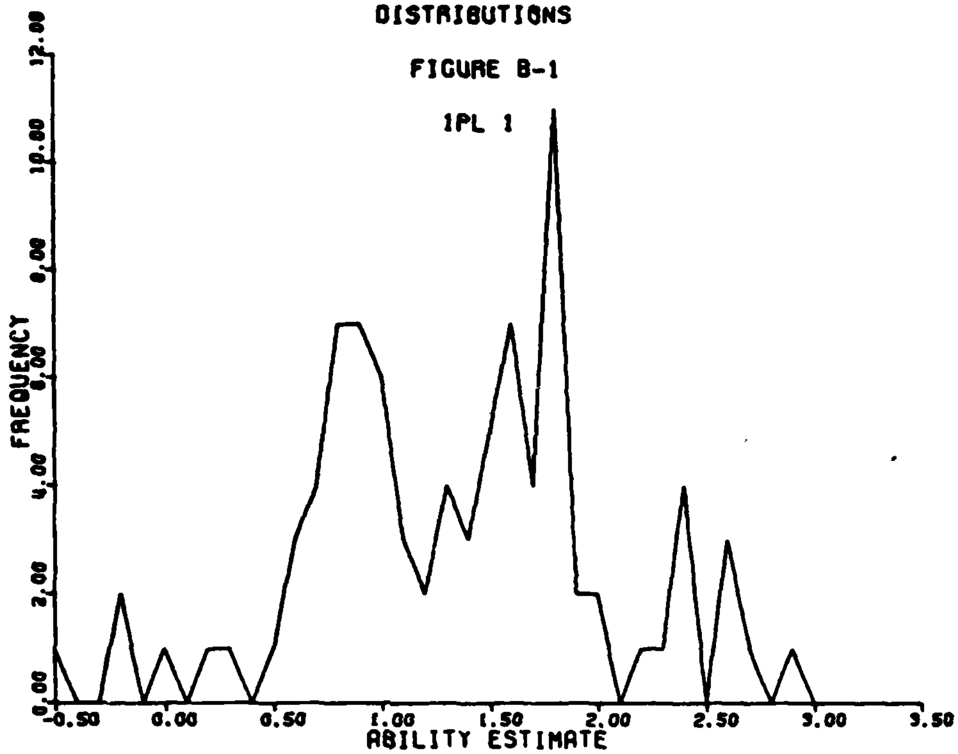
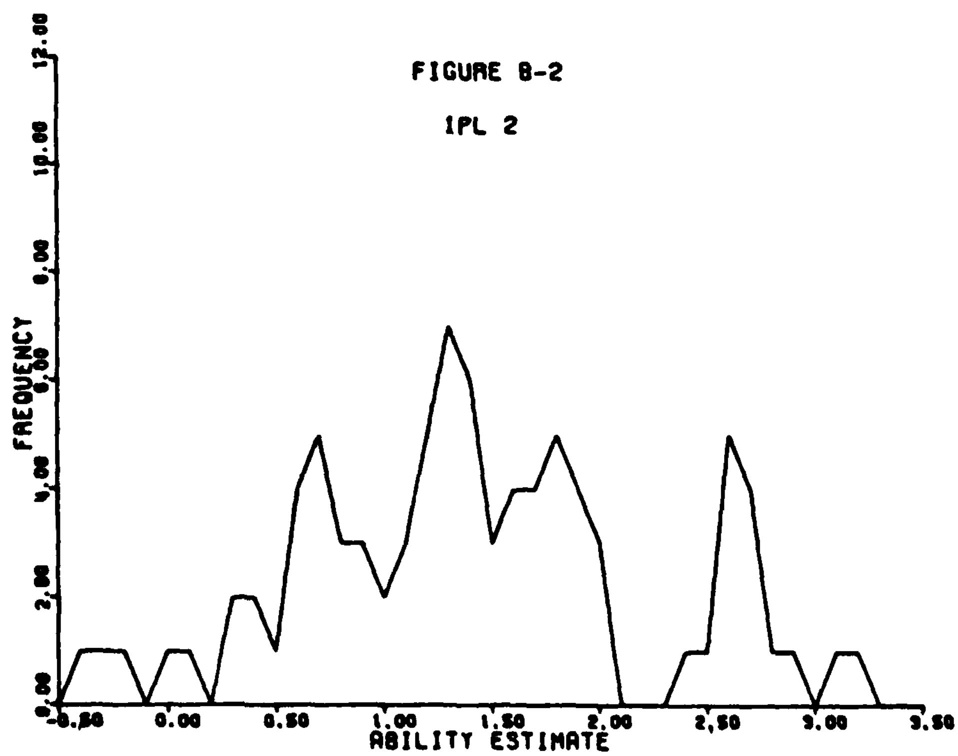


FIGURE B-2

IPL 2



ABILITY ESTIMATE FREQUENCY

DISTRIBUTIONS

FIGURE B-3

3PL 1

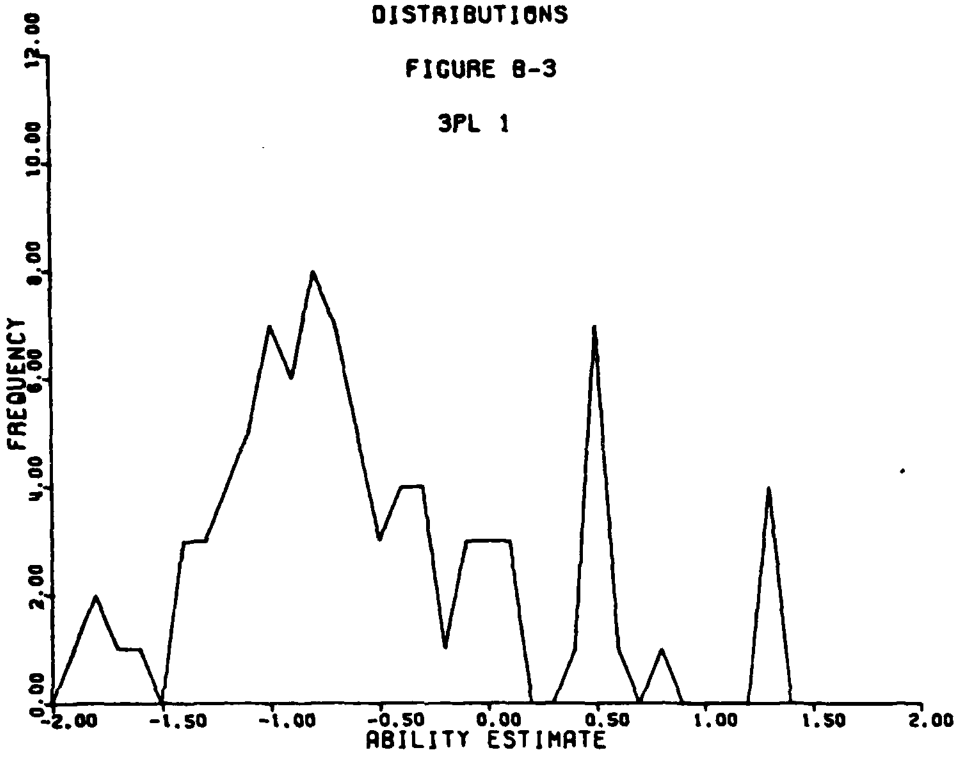
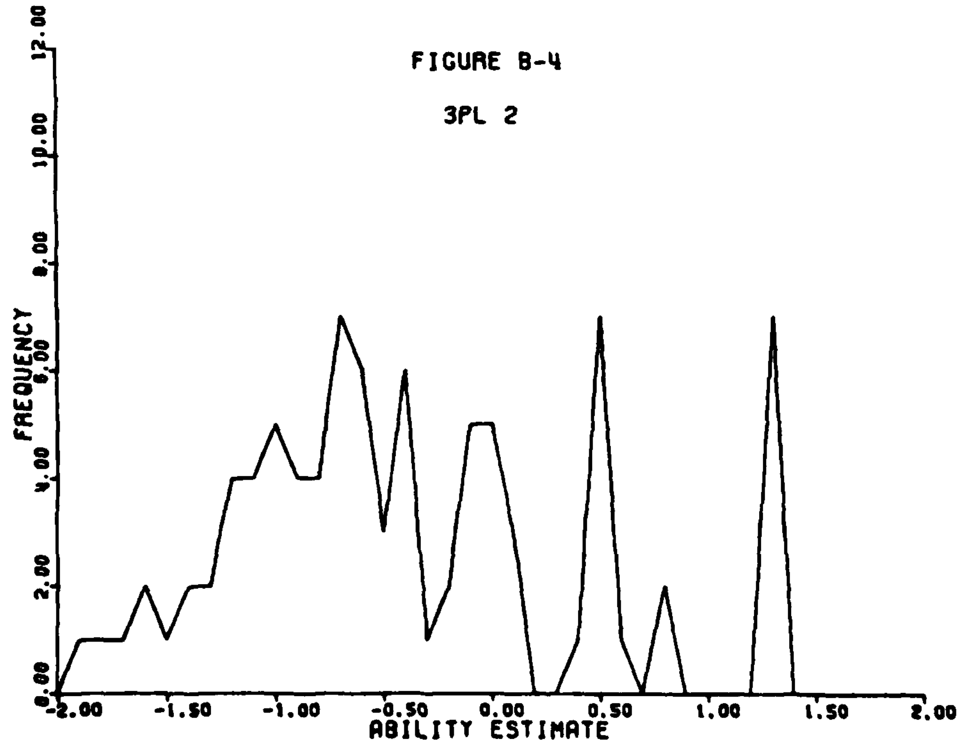


FIGURE B-4

3PL 2



APPENDIX C

Attitude Survey Administered After Each Tailored Testing Session

Please circle the response to each statement below which most nearly reflects your feelings or attitude.

1. During the test I was worried about how well I was doing.

   strongly agree | agree | neutral | disagree | strongly disagree

2. I felt less time pressure while taking this computerized test than while taking conventional tests.

   strongly agree | agree | neutral | disagree | strongly disagree

3. I felt that many of the items were too difficult for me.

   strongly disagree | disagree | neutral | agree | strongly agree

4. The computer terminal made me feel that I had to answer the items as quickly as possible.

   strongly agree | agree | neutral | disagree | strongly disagree

5. I didn't care very much about how well I did on the test.

   strongly disagree | disagree | neutral | agree | strongly agree

6. I think I did well on the test compared to other people.

   strongly agree | agree | neutral | disagree | strongly disagree

7. I felt that my performance on this test reflected my true knowledge of A140.

   strongly disagree | disagree | neutral | agree | strongly agree

8. My eyes were uncomfortable when viewing the screen.

   strongly agree | agree | neutral | disagree | strongly disagree

9. I felt that most of the items on this test were too easy.

   strongly disagree | disagree | neutral | agree | strongly agree

10. I was nervous about coming here to take this test.

strongly agree	agree	neutral	disagree	strongly disagree

11. The pace of the computer was so slow that it made me impatient.

strongly disagree	disagree	neutral	agree	strongly agree

12. I feel that I did as well on this test as on other tests I've taken.

strongly agree	agree	neutral	disagree	strongly disagree

13. The computer terminal made me nervous.

strongly agree	agree	neutral	disagree	strongly disagree

14. I felt confident that I did well on the test.

strongly disagree	disagree	neutral	agree	strongly agree

15. I felt considerable stress while taking the test.

strongly disagree	disagree	neutral	agree	strongly agree

16. It was easy to read the words and questions on the screen.

strongly agree	agree	neutral	disagree	strongly disagree

17. I felt that the test did a good job of measuring my ability in A140.

strongly agree	agree	neutral	disagree	strongly disagree

18. I think I could have done better on the test if I had tried harder.

strongly disagree	disagree	neutral	agree	strongly agree

19. I was careful to try to select the best answer to each question.

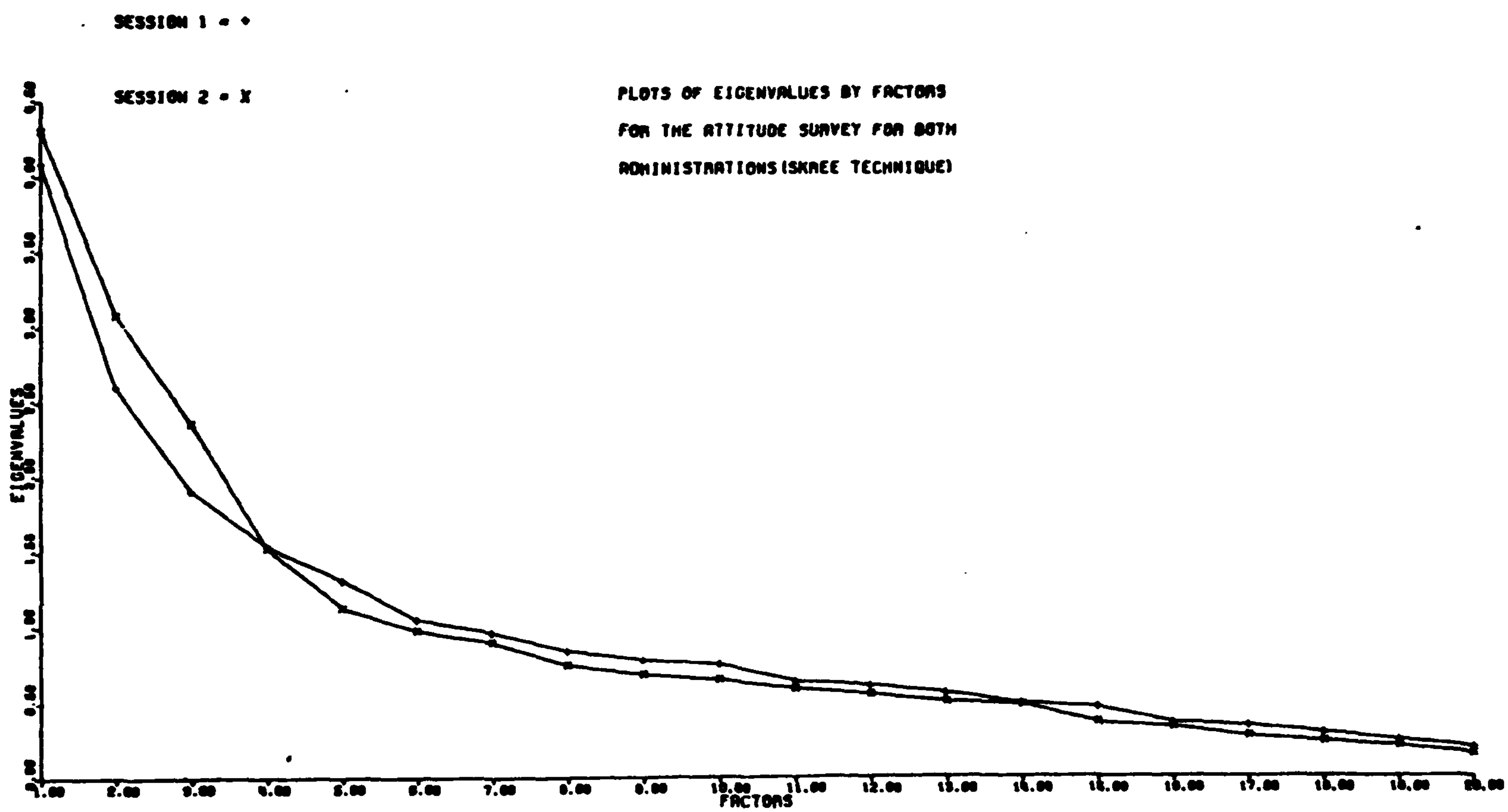
strongly disagree	disagree	neutral	agree	strongly agree

20. I tried to finish the test quickly just to receive my extra credit points.

strongly agree	agree	neutral	disagree	strongly disagree

APPENDIX D

PLOTS OF EIGENVALUES BY FACTORS
FOR THE ATTITUDE SURVEY FOR BOTH
ADMINISTRATIONS (SKREE TECHNIQUE)



DISTRIBUTION LIST

Navy

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Chief of Naval Education and Training
Liason Office
Air Force Human Resource Laboratory
Flying Training Division
WILLIAMS AFB, AZ 85224

1 COMNAVMILPERSCOM (N-6C)
Dept. of Navy
Washington, DC 20370

1 Deputy Assistant Secretary of the Navy
(Manpower)
Office of the Assistant Secretary of
the Navy (Manpower, Reserve Affairs,
and Logistics)
Washington, DC 20350

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

Navy

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. George Moeller
Head, Human Factors Dept.
Naval Submarine Medical Research Lab
Groton, CN 06340

1 Dr. William Moonan
Code 303
Navy Personnel R&D Center
San Diego, CA 92152

1 Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

1 Library
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92138

1 Naval Medical R&D Command
Code 44
National Naval Medical Center
Bethesda, MD 20014

Navy

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

5 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

1 Director, Navy Personnel R&D Center
Washington Liason Office
Building 200, 2N
Washington Navy Yard, DC 20374

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

Navy

1 Special Asst. for Education and
Training (OP-01E)
Rm. 2705 Arlington Annex
Washington, DC 20370

1 Office of the Chief of Naval Operations
Research, Development, and Studies Branc
(OP-102)
Washington, DC 20350

1 Long-Range Manpower, Personnel, and
Training Planning Branch (OP-110)
Room G828
Arlington Annex
Washington, DC 20350

1 Captain Donald F. Parker, USN
Commanding Officer
Navy Personnel R&D Center
San Diego, CA 92152

1 LT Frank C. Petho, MSC, USN (Ph.D)
Code L51
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Director, Research & Analysis Division
Plans and Policy Department
Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 HQ USAREUE & 7th Army
ODCSOPS
USAAREUE Director of GED
APO New York 09403

1 Col Gary W. Bloedorn
US Army TRADOC Systems Analysis Activity
Attn: ATAA-TH
WSMR, NM 88002

1 DR. RALPH DUSEK
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Army

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Frederick Steinheiser
U. S. Army Reserch Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

CoastGuard

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 Air University Library
AUL/LSE 76/443
Maxwell AFB, AL 36112

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Philip De Leo
AFHRL/TT
Lowry AFB, CO 80230

1 Dr. Genevieve Haddad
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Frank Schufletowski
U.S. Air Force
ATC/XPTD
Randolph AFB, TX 78148

1 Jack A. Thorpe, Maj., USAF
Naval War College
Providence, RI 02846

1 Brian K. Waters, Lt Col, USAF
Air War College (EDV)
Maxwell AFB, AL 36112

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Headquarters, U. S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Special Assistant for Marine Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

Other DoD

12 Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, VA 22314
Attn: TC

1 Dr. Dexter Fletcher
ADVANCED RESEARCH PROJECTS AGENCY
1400 WILSON BLVD.
ARLINGTON, VA 22209

1 Director, Research and Data
OASD(MRA&L)
3B919, The Pentagon
Washington, DC 20301

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 MAJOR Wayne Sellman, USAF
Office of the Assistant Secretary
of Defense (MRA&L)
3B930 The Pentagon
Washington, DC 20301

Civil Govt

1 Dr. Susan Chipman
Learning and Development
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Lorraine D. Eyde
Personnel R&D Center
Office of Personnel Management of USA
1900 EStreet NW
Washington, D.C. 20415

1 Jerry Lehnus
REGIONAL PSYCHOLOGIST
U.S. Office of Personnel Management
230 S. DEARBORN STREET
CHICAGO, IL 60604

1 Dr. Joseph I. Lipson
SEDR W-638
National Science Foundation
Washington, DC 20550

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Personnel R&D Center
Office of Personnel Managment
1900 E Street NW
Washington, DC 20415

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
600 Jay Street
Sacramento, CA 95819

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

Non Govt

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

Non Govt

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ross Green
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

Non Govt

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Douglas H. Jones
Rm T-255
Educational Testing Service
Princeton, NJ 08450

3 Journal Supplement Abstract Service
American Psychological Association
1200 17th Street N.W.
Washington, DC 20036

1 Dr. Arnold F. Kanarick
Honeywell, Inc.
Honeywell Plaza MN12-3166
Minneapolis, MN 55408

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 Dr. Michael Levine
210 Education Building
University of Illinois
Champaign, IL 61820

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat
Groningen
NETHERLANDS

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

Non Govt

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. John R. Olsen
8740 Scenic Highway
Pensacola, FL 32504

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

Non Govt

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

Non Govt

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Kathryn T. Spoehr
Department of Psychology
Brown University
Providence, RI 02912

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 Dr. David Stone
ED 236
SUNY, Albany
Albany, NY 12222

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

Non Govt

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Douglas Towne
Univ. of So. California
Behavioral Technology Labs
1845 S. Elena Ave.
Redondo Beach, CA 90277

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Bureau of Social SCience Research
1990 M Street, N. W.
Washington, DC 20036

1 Dr. John Wannous
Department of Management
Michigan University
East Lansing, MI 48824

Non Govt

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2